Towards Universal Speech Recognition

Zhirong Wang, Umut Topkara, Tanja Schultz, Alex Waibel
Interactive Systems Laboratories
Carnegie Mellon University,
Pittsburgh, PA, 15213
Email: {zhirong, tanja, ahw}@cs.cmu.edu, utopkara@cs.purdue.edu

Abstract

The increasing interest in multilingual applications like speech-to-speech translation systems is accompanied by the need for speech recognition front-ends in many languages that can also handle multiple input languages at the same time. In this paper we describe a universal speech recognition system that fulfills such needs. It is trained by sharing speech and text data across languages and thus reduces the number of parameters and overhead significantly at the cost of only slight accuracy loss. The final recognizer eases the burden of maintaining several monolingual engines, makes dedicated language identification obsolete and allows for code-switching within an utterance. To achieve these goals we developed new methods for constructing multilingual acoustic models and multilingual n-gram language models.

Keywords: Multilingual acoustic modeling, data-driven, IPA, Multilingual n-gram language modeling

1. Introduction

With the appearance of low-cost commercial speech processing software, spoken language applications are transferred ever more rapidly into practical use. This comes with a growing interest in expanding the reach of speech and language systems to international markets and consumers worldwide. As a consequence, today's multilingual applications such as speech-to-speech translation systems inquire for speech recognizer front-ends which can not only handle input from many languages, but also switch between those languages instantly.

So far, the majority of speech recognizers can only handle one language at a time. For the multilingual speech-to-speech translation system Verbmobil for example, the problem of handling several languages was solved using a dedicated language identification (LID) module that first determined the spoken language and then triggered to the appropriate monolingual recognition system [1]. However, fast and reliable LID is still a challenging task and triggering to language specific recognizers requires time and the storage of each recognizer in the memory separately. Moreover, in such a setup, switching to another language is only possible at the beginning of a new utterance. Most work that has been done on handling multiple languages at a time was focused on building multilingual acoustic models by sharing data across languages [2], only few publications deal with multilingual language models [3,4] and the combination of both into one engine [5].

In this paper we describe the development and investigation of a *universal* or *multilingual* speech recognition system. The acoustic and language model of the recognizer is trained by sharing speech and text data across languages. It consists of a multilingual acoustic model that covers the sounds of all languages in question, a dictionary combining the words of these languages and a language model that allows for code-switching, i.e. switching the input language within an utterance. Such a universal speech recognizer has several benefits: (1) since it is one single engine with multilingual sources it is much easier to maintain than several monolingual engines, (2) it is suitable for multilingual applications without the need for (2a) performing language identification to trigger to the appropriate engine and without the need for (2b) loading and switching between those engines, (3) it enables code-switching, and (4) it allows to counterbalance data sparseness of some languages by sharing data across all languages.

Our investigation in this paper focused on two languages: English and German. We observed significant differences in recognition performance that are partially due to a higher acoustic confusability (e.g.,

English), and a larger number of compounds and richer inflection (e.g., German). Such distinctions put a different burden on acoustic modeling vs. language modeling. We are investigating the recognition performance of these two languages in the multilingual setting. The paper is organized as follows. First section 2 describes the used data and discusses various approaches for merging speech phonemes across languages. Then section 3 investigates the multilingual n-gram language modeling issues. Section 4 presents the experimental results of acoustic and language modeling. Section 5 gives a brief summary and conclusions.

2. Multilingual Acoustic Modeling

In our work a single bilingual recognizer was built with a large-size vocabulary that contains the words from both languages to reduce the computation load. For the acoustic models, we defined a global speech unit set by merging phones from different languages. This idea is based on the belief that some phones across different languages may be similar enough to be equated. These language independent phones allow the data and model sharing of various languages to reduce the complexity and number of parameters in the bilingual LVCSR system.

2.1 Speech data

For the training data, we have about 60 hours German speech data (GSST) and 40 hours of English speech data (ESST) from Verbmobil-II project; these data features spontaneous speech on a limited domain under relatively clean acoustic conditions. Since the amount of English speech data is much less than that of German data, we added 15 hours of English Broadcast News (BN) data to the training database. The BN data consists of clean, read speech data from a very large domain.

Database	German	English
Training data	60h(Spontaneous)	40h(Spontaneous) 15h(Read)
Vocabulary	10K	40K
Testing data	61 minutes 30 speakers 744 turns	58 minutes 56 speakers 290 turns

Table 1 Data

For the testing data, the final German evaluation was carried out on the GSST eval00 test set; the English one was carried out on the part of BN 98 evaluation data set. Table 1 shows the details of our data set.

2.2 Knowledge-based model sharing

The idea of knowledge-based model sharing in our research is based on the assumption that the articulatory representations of phonemes are so similar across languages that phonemes can be considered as units that are independent from the underlying language. This idea was first proposed by the International Phonetic Association [6]. In this method the similarities of sounds are documented and classified based on phonetic knowledge. Sounds of different languages, which are represented by the same IPA symbols, share one common unit. The main motivation for sharing common units across different languages is to make better use of available data in training Gaussian codebooks, when features of the training data from two languages are located closely in the acoustic space, they are used in training one common codebook.

After the mapping, there are several ways to combine these IPA units from different languages into one. One way is to preserve the language information for each phoneme, so that each language-specific phoneme is trained solely with data from its own language; the second way is to mix these phonemes together, those phonemes of different languages which belong to the same IPA units are sharing data from different languages during training, the language information is not preserved anymore. These are not the best ways to mix the IPA phones together according to the research of globalPhone project [2]. From the globalPhone project, we know that previous mentioned approaches are outperformed by the tag method if used for recognizing one of the training languages. So here we used the tag method to carry out our experiments. In this method, each phoneme receives a language tag attached in order to preserve the information about the language the phoneme belongs to. During the training, the Gaussian components are shared across languages, but the mixture weights are kept separately for different languages.

The main advantage of IPA-based approach is that it is a simple way of getting multilingual models, and it is easily be applied to many different languages. The disadvantage is that the IPA method does not consider the spectral properties and the statistical similarities of the phone models.

2.3 Data-driven model sharing

The basis of the data-driven methodology is a number of iteratively conducted bottom-up clustering steps. The

clustering procedure is initialized with language-specific phoneme models, the strategy is to select and merge those phonemes that correspond to the two most similar speech units iteratively. The measurement of similarity between two phone models was defined before the clustering. This method considers the spectral properties and the statistical similarities of the phone models, but it is hard to transfer these clusters to new languages. We tried this method on both context independent and dependent phones.

2.3.1 Context independent modeling

For the context independent modeling using data-driven method, we trained a context-independent system with phones from both languages, in which each phone shares the same Gaussian component but has its own mixture weights. In this way, we defined the similarity between two phone models as the distance between their mixture weights. We used Euclidean distance as the distance measurement method. At each clustering step, the most similar pair of clusters is merged to a new cluster. Because the estimation of the new phone models of the merged cluster is difficult to achieve, the distance of two clusters is always computed with the original phone models that are the basic elements of one cluster, the distance between two clusters is determined with the furthest neighbor criterion.

The clustering process continues until all calculated clusters distance are higher than a pre-defined distance threshold, or we can stop the clustering when a specified number of clusters are achieved. After the clustering procedure, we defined each cluster as a new phone model for the bilingual system. In this experiment, in order to compare the data-driven method with knowledge-based method, we specified the number of clusters to stop the iteration. In this way, we got the same number of phones from data-driven method as from the knowledge-based method.

Table 2 shows merged results from IPA and data-driven on context independent modeling method. The table indicates that the IPA-based and data-driven method seem to agree on merging consonants while vowels are more diverse.

2.3.2 Context dependent modeling

For the previous two methods, we worked only on context independent acoustic models. Actually the left and right contexts are two very important contribution factors that affect the realization of a phone especially in spontaneous speech. From the experience of language dependent case wider contexts increase recognition performance significantly, we want to investigate whether such improvement extend to the multilingual setting.

Phones combined by both IPA and Data-driven method

English	German
n (N)	n (N)
h (HH)	h (H)
z (Z)	z (Z)
ï (IY)	ï (IE)
f (F)	f (F)
S (S)	S (S)
ŋ (NG)	Ŋ (NG)
ʃ (SH)	ʃ (SCH)
b (B)	b (B)
m (M)	m (M)

Phones combined only by IPA method

English	German
v (V)	v (V)
ʧ (CH)	ʧ (TSCH)
ę (AX)	Ę (E2)
g (G)	g (G)
l (L)	l (L)
J (Y)	j (J)
Ɛ (EH)	Ɛ (AEH)
d (D)	d (D)
U (UW)	U (U)

Phones combined only by Data-driven method

English	German
ɵ (T)	t (T)
ɔ (AA)	ä (AH)
I (IH)	e (E)
ai (AY)	ai (AI)
Ƙ (K)	k (K)
ɯ (W)	ü (UH)
^(AH)	a (A)
au (AW)	ä (AH)
Oɪ (OW)	au (AU)

Not combined by any method

English	German
Ƥ (P)	-
Ď (DH)	-
Ŕ (R)	-
ʤ (JH)	-
Ť (TH)	-
ɗ (DX)	-
ʒ (ZH)	-
Æ (AE)	-
ɾ̃ (AXR)	-
>(AO)	-
ɾ̌ (ER)	-
ei (EY)	-
Ŭ (UH)	-
Ĭ (IX)	-
İ (OY)	-
-	Ŋ (ANG)
-	v (OE)
-	i (I)
-	Ç (CH)
-	o (O)
-	p (P)
-	r (R)
-	Ÿ (UEHR)
-	x (X)
-	ê (HER)
-	î (IHR)
-	ô (OR)
-	â (AHR)
-	ts (TS)
-	õ (OHR)
-	ë (EH)
-	Ǝ (ER2)
-	ö (OH)
-	û (UHR)
-	eu (EU)
-	á (AR)
-	é (ER)
-	ÿ (UEH)
-	y (UE)
-	í (IR)
-	ɐ (OEH)
-	É (AEHR)
-	ú (UR)

Table 2 Phones merging information

The first step towards getting context dependent phone models for multilingual speech units is to collect all the contexts that can be modeled with the given task. Here we limited the maximum context width to 1 to both sides, and at this time we didn't allow cross-word

contexts that go from one word into the neighboring word. These phones with left and right contexts are called triphones, they are powerful because they capture the most important coarticulatory effects in spoken language, and they are generally much more consistent than the context independent phone models. The triphones are collected from all the training data. During the collection, the transcription text of every utterance is examined, optional silences can be inserted between words and optional alternative pronunciation variants can be allowed.

We will easily get a lot of different triphones, when the training corpus is large and the dictionary contains many variants. And most likely we wouldn't have enough training examples to estimate the acoustic model for every triphone. So we have to limit the triphone types to be included in our bilingual phone set. Figure 1 shows the triphone type/token relation in our training corpus; from this graph we chose the 400 most frequent triphones plus the context independent phone models as our new bilingual phone set.

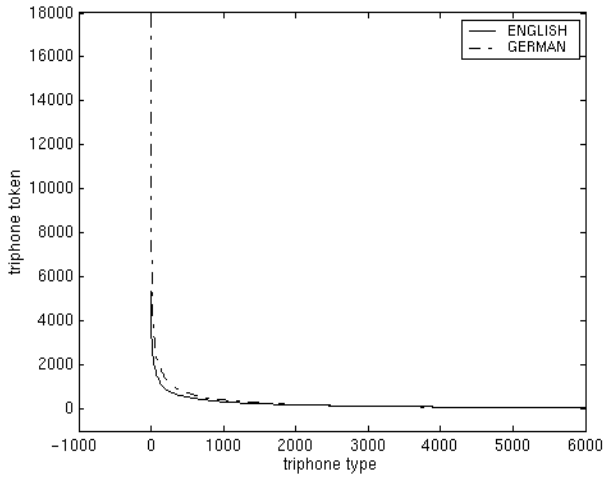

Figure 1 Difference in triphone occurrence between English and German

Figure 2 shows the coverage of the triphones between testing and training data. The x-axis shows the number of triphone types from the training corpus, and the y-axis shows the number of triphone tokens from the testing corpus. From the graph we can see that for the same speaking style, since the English has less variation of triphones, ESST testing data is covered by the training data much better than that of GSST data. While comparing the ESST with BN data, we can see that the different speaking styles also have a strong influence on triphone coverage.

After we got the bilingual speech units using different approaches, the Janus recognition Toolkit was used to train the fully continuous HMM systems. For each system, a mixture of 32 Gaussian components is assigned to each state of a polyphone. The Gaussians are on 13 Mel-scale cepstral coefficients with first and second order derivatives, power and zero crossing rate. Incorporated into our continuous HMM systems are techniques such as linear discriminate analysis (LDA) for feature space dimension reduction, vocal tract length normalization for speaker normalization, cepstral mean normalization for channel normalization and wide-context phone modeling. The recognition results of various systems are presented in section 4. At this time, we did the English tests only on BN data, latter we will do these experiments on ESST data.

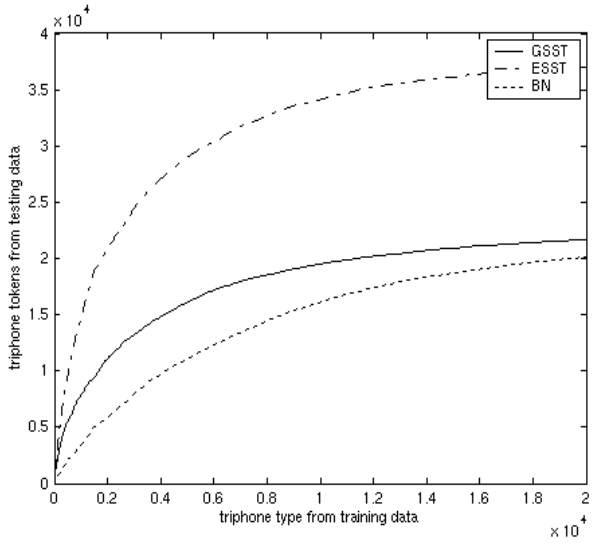

Figure 2 triphone coverage

3. Multilingual n-gram Language Modeling

The promise of our multilingual decoder is being able to recognize utterances from several languages under a single process. Building such a system requires a multilingual acoustic model and a multilingual language model (LM). We define a multilingual LM as a single stochastic model that captures the linguistic behavior of speech that has mixed usage of several languages. This can be in a conversation that occurs between parties speaking different languages, or a dictation monologue where the speaker is bilingual. Also switching between languages is allowed at arbitrary positions in sentences. This is especially important when a speaker can speak more than one language, or when some concepts are referred to with their names in one of the languages as the conversation develops.

The vocabulary of a multilingual LM has to satisfy some requirements for the decoder to work correctly.

First, the multilingual vocabulary has to be a superset of the vocabularies of languages covered. Also, each entry in the multilingual vocabulary has to be tagged with the language it belongs, so as to distinguish between the homonyms among the covered languages.

In order to compare different multilingual language modeling approaches we used one of our multilingual acoustic models, and ran experiments on monolingual test cases by plugging in different LMs. The details of German test data for the following experiments can be found in table 1. For the English test data, we used a different one from what we described in table 1. This English test set was recorded in our lab, contains only 198 turns from 2 speakers, the speaking style is similar to BN data. Table 3 below summarizes our results in this stage:

LM type	German	English
Best possible	27.8	14.8
Experiment 1	44.0	15.5
Experiment 2	39.8	57.8
Experiment 3	35.6	16.1
Experiment 4	31.0	17.0

Table 3: Multilingual Language Modeling results [WER %]

Differences in linguistic nature of the languages and available data for them are common properties of multilingual data collections and they can complicate multilingual language modeling. In our case, the English and German corpora are unbalanced in their size with a ratio of 218 to 1 favoring English side. Respectively, the English vocabulary we use is 4 times larger than the German vocabulary.

The first row in Table 3 shows the performances of two decoders that have monolingual LMs trained separately on our two corpora, and are the best possible performance that can be achieved by a decoder with a multilingual LM in this setting.

The first approach we have tried is to concatenate corpora in hand and compute the probabilities for a multilingual LM from the resulting corpus. When we plainly concatenated English and German corpora in *Experiment 1*, performance on German recognition becomes extremely poor. Since the German text constitutes a relatively tiny portion of the combined corpus, and German n-grams are assigned smaller probabilities compared to English n-grams. This causes German words to be incorrectly recognized as English words in decoding, especially when the LM backs off to 1-grams. Same situation happens for English words, when acoustically confusable German words with high probability exist in the LM.

One of the most common methods used to combine statistical data obtained from different sources of information is to use linear interpolation. We created the multilingual LM in *Experiment 2* by interpolating two monolingual LMs with equal weight. Linear interpolation performs poorly on both English and German recognition. The overall probability distribution functions for two languages are different, that is most of the n-gram probabilities in the monolingual German LM are higher than most of n-gram probabilities in the monolingual English model. When these LMs are interpolated with equal weights, German n-grams dominate English n-grams with their high probabilities and the decoder incorrectly hypothesizes German words for English utterances.

We contribute to these experiments by a new interpolation scheme. In this scheme, we try to balance the probability distribution functions of two languages, rather than balancing the probability mass assigned to them. Our scheme assigns similar probabilities to two n-grams obtained from different corpora if they are at similar positions with respect to rest of n-grams obtained from their respective corpora. To show the concept, in our experiments we used the frequency ranks of n-grams to judge on their similarity. It is defined as the position of an n-gram from top when n-grams are sorted with respect to their frequencies.

In *Experiment3* we assigned German 1-gram frequencies to English 1-grams frequencies that have the same frequency rank. Then we incremented higher order German n-gram frequencies with the same increase ratio of their lower order n-gram frequencies from left. The resulting multilingual LM performs comparably better than other approaches. Then, in *Experiment 4*, we directly assigned higher order German n-gram frequencies from corresponding English n-gram frequencies. Although still far away from achieving monolingual recognition rates, these two methods both outperform traditional methods. Good performance of these LMs show that balancing the probability distribution among individual n-grams brings important performance gains to multilingual language modeling.

4. Experimental Results

We tested the usefulness of our modeling approaches by comparing the recognition performance, which is achieved by the resulting systems from different acoustic and language modeling methods. All the English experiments were tested on the BN evaluation set with 290 turns from 56 speakers, while all the German experiments were tested on the Verbmobil-II eval00 test set with 30 speakers (see table 1 for the detail).

Here is the information of the baseline systems. For English, we are using the Broadcast News speech

recognizer as the baseline system; this system achieves a first pass WER of 19.0% on all F-conditions of BN task, and 18.2% on our testing data set. For German, the Verbmobil system was used, and the WER on the eval00 testing data is 25.5%. To be comparable to these baseline systems, we used the same setup as the baseline system to build the bilingual system; only the set of phone models and the language model are different.

Table 4 shows the word error rate from various systems. Column 1 indicates whether the acoustic model is from IPA-based method or data-driven method that were described in section 3. DD_CI means context independent models from data-driven method, and DD_CD means the context dependent models from data-driven method. Column 2 indicates whether the LM is a monolingual LM or a bilingual LM. For the bilingual LM we used the new-scaled bilingual language model, which was described in section 3.

Compared to the baseline systems with using the same monolingual language model, both IPA models and the context independent models from data-driven method are nearly as good as the language-dependent models. The decrease in recognition rate is about 1% with 150K densities instead of 270K densities in the language-dependent case. The data-driven approach is able to detect and exploit the acoustic phonetic similarities across the phones of different languages; from this table we can see that the context independent models from data-driven method outperforms the IPA method in German, but not in English. This may due to the differences in the quality and recording conditions of BN and GSST corpora. For the context dependent models from data-driven method, it does help to improve the performance of German, but hurts the English recognition; we attribute this to the poorer coverage of English triphones in testing data than that in German testing data.

AMs	LMs	English(%)	German(%)
Baseline		18.2	25.5
IPA	Mono	18.5	26.5
IPA	Bilingual	20.6	29.2
DD_CI	Mono	19.2	26.2
DD_CI	Bilingual	21.1	28.5
DD_CD	Mono	22.7	25.6
DD_CD	Bilingual	24.4	27.8

Table 4 Recognition results (WER)

On the other hand, using the bilingual language model results the degradation of performance by an average of 2.1%(1.7%~2.7%). Nearly all of this loss is due to false transitions from one language to the other language in the middle of a hypothesis. Main actor in this performance loss is the acoustic confusability between words in two languages. German utterances suffer more, because its n-grams have low scores due to morphological richness of German. On the English side, frequent occurrences of less likely words, a characteristic of the test cases, causes false language switching. Table 5 shows the language false language switching rates from our experiments:

For the 290 English sentences, there are 26 hypotheses contain German words, the mixing rate is about 9.8%, while for German sentences, the mixing rate is about 15.0%.

Language	Hypotheses in one language	Hypotheses with mixed languages	Mixing rate
English	264 turns	26 turns	9.8%
German	659 turns	115 turns	15.0%

Table 5 Language mixing rate

5. Summary and Future Work

In this paper, we addressed language dependent and independent acoustic modeling and language modeling for multilingual speech recognition. The multilingual engine allows code-switching, that is switching of the language within one sentence and recognition of more than one language without changing recognizer. The experiments show that the bilingual system can achieve comparable performance with the monolingual systems and at the same time reduce a huge number of parameters.

6. References

[1] A. Waibel, H. Soltau, T. Schultz, T. Schaaf, and F. Metze. Multilingual Speech Recognition. In *Verbmobil: Foundations of Speech-to-Speech Translation*, W. Wahlster (Ed.), Springer Verlag, 2000.

[2] T. Schultz and A. Waibel. Language Independent and Language Adaptive Acoustic Modeling. In *Speech Communication*, Vol 35, Issue 1-2, pp 31-51, August 2001.

[3] S. Harbeck, E. Nöth, H. Niemann. Multilingual Speech Recognition. In SQEL, 2nd Workshop on Multi-Lingual Information Retrieval Dialogs, Plzeň, Czech Republic, April 1997.

[4] F. Weng, H. Bratt, L. Neumeyer, A. Stolke. A Study of Multilingual Speech Recognition. In *EUROSPEECH*, Rhodos, Greece, September 1997.

[5] T. Ward, S. Roukos, C. Neti, M. Epstein, S. Dharanipragada. Towards Speech Understandig across Multiple Languages. In *ICSLP*, Sydney, Australia, November 1998.

[6] IPA, (1993). The International Phonetic Association (revised to 1993) IPA Chart. *Journal of the International Phonetic Association* 23, 1993.